REMARKS

Upon entry of this response claims 1-5, 7, 11-21, 24-25, 28-37, 40-41 and 43-45 are pending, and of these claims 1, 3, 28, and 43 are independent.

Applicants have amended the specification to include specific disclosure incorporated by reference in the original application according to 37 CFR §1.57(f). In particular, Applicants have amended the description to include subject matter described in U.S. Provisional Patent Application Serial No. 60/398,958 which was properly incorporated by reference in the present application in paragraph [0120]. The subject matter pertains to an example of iteratively fitting intensity values that correspond to probes on a probe array with models of genomic structure to indicate the presence of alternative splice variants. Applicants have attached a copy of the '958 application as filed to the present response and direct the Examiners attention to the disclosure on pages 10-14, and 15-18 for support of the present amendments.

Applicants have also amended claims 1, 3, 28, and 43 to add clarity to the user selection that comprises one or more probe-set identifiers and one or more intensity values (support may be found in paragraph [0119]). In addition, Applicants have amended each of the aforementioned claims the add the limitations of each of the probe-sets comprising one or more probes and intensity data detected from the probes, also for the purpose of clarity (support may be found in paragraphs [0080] and [0063]).

Applicants respectfully assert that no new matter is presented by these amendments. Applicants respectfully request entry of the same.

Reply to Claim Rejections – 35 U.S.C. §101

Claims 1, 3-5, 7, 11-21, 24-25, 28-37, and 45 are rejected under 35 U.S.C. §101.

Upon further review of the Examiners' remarks in the office actions mailed 2/11/2005, and 10/6/2005, as well as the subject matter of a personal interview conducted in co-pending application serial no. 10/065,868. It is the Applicants understanding that the Examiner feels that a specific utility limitation must be recited in the claims, such as the example of a diagnostic limitation of a probe array provided by the Examiner (see office action of 2/11/2005 stating that a diagnostic correlation with a disease is not recited in the claim).

Applicants respectfully disagree with such a position and respectfully assert that there is no requirement in the law for such specific utility limitations in the claims in order to satisfy the utility requirements. Rather, the current state of the law indicates that a practical utility of the claimed invention that is specific and substantial must either be obvious to one of ordinary skill in the related art or be asserted in the disclosure of the application. For example, the case of *Cross v. Iizuka* indicates that a claim meets the requirements for utility when there is evidence of practical utility, even though the claim does not recite any particular utility (See *Cross v. Iizuka*, 753 F.2d 1040, 224 USPQ 739 (Fed. Cir. 1985)). The *Cross v. Iizuka* case further indicates that the requirement of substantial or practical utility is met if the utility is either obvious or discovered and disclosed in the application.

Applicants believe that the Examiners' position inappropriately attempts to place limitations upon the claims that are not supported by the current state of the law. Specifically, the Examiner is attempting to improperly limit the patentable subject matter

of the claimed invention to one specific use. Applicants respectfully assert that a claimed invention is not required to be limited to one specific use, even when more than one specific use is disclosed. Applicants respectfully direct the Examiners attention to the case of *Ex parte* Lanham that involved an invention of a compound and process for making it that had two disclosed utilities, (1) a solvent and softening agent, and (2) an intermediate. The Board determined that the disclosure of a single utility was sufficient to meet the requirements for utility. (See *Ex parte* Lanham, 121 USPQ 223 (Pat. Off. Bd. App. 1958))

Applicants respectfully point the Examiner to the recently published (OG Notices: 22 November 2005) "Interim Guidelines for Examination of Patent Applications for Patent Subject Matter Eligibility" document. The Interim Guidelines provide general guidance for the Examiner with respect to the subject of improperly placing limitations on subject matter that may be patented, where the placement of said limitations are not supported in the law. Applicants believe that such guidance has particular relevance to the present question of whether a claim must recite specific utility limitations. For example, the paragraph from Section IV (A) refers to this:

> The plain and unambiguous meaning of section 101 is that any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may be patented if it meets the requirements for patentability set forth in Title 35, such as those found in sections 102, 103, and 112. The use of the expansive term "any" in section 101 represents Congress's intent not to place any restrictions on the subject matter for which a patent may be obtained beyond those specifically recited in section 101 and the other parts of Title 35 . . . Thus, it is improper to read into section 101 limitations as to the subject matter that may be patented where the legislative history does not indicate that Congress clearly intended such limitations.
> Alappat, 33 F.3d at 1542, 31 USPQ2d at 1556

Applicants respectfully assert that the Examiners position of requiring specific utility limitations in the claims contradicts the guidance set forth above. For example, the rulings in *Cross v. Iizuka* and *Ex parte* Lanham clearly indicate that a claim meets the utility requirements when at least one utility for a practical application is asserted in the disclosure. Thus, the Examiners position of requiring the claims to recite a specific utility in the limitations is improper.

Applicants would further like to point the Examiner to the guidance of the MPEP that pertain to determinations of asserted utility that are specific and substantial. MPEP §2107 (C)(1) states:

> (1) Where the ***asserted utility is not specific or substantial***, a prima facie showing must establish that it is more likely than not that a person of ordinary skill in the art would not consider that any utility ***asserted*** by the applicant would be specific and substantial. The prima facie showing must contain the following elements:
>
> (i) An explanation that clearly sets forth the reasoning used in concluding that the ***asserted*** utility for the claimed invention is not both specific and substantial ***nor well-established***;
>
> (ii) Support for factual findings relied upon in reaching this conclusion; and
>
> (iii) ***An evaluation of all relevant evidence of record, including utilities taught in the closest prior art.***

The guidance above clearly directs the Examiner to consider all asserted utility. Applicants also point out that said guidance does not give any suggestion that claims must recite limitations to specific utility. Instead, the guidance from the MPEP directs the Examiner to consider all relevant evidence of record as well as utilities taught in the closest prior art for an asserted utility.

As will be described in greater detail below, Applicants respectfully assert that a practical utility that is both specific and substantial for the presently claimed invention is

both asserted in the description of the application, and obvious to one of ordinary skill in the art. Applicants acknowledge that the claimed invention as a whole must have an asserted use as set forth in section 101, and further discussed in *Raytheon Co. v. Roper Corp.* (See Raytheon Co. v. Roper Corp., 724 F.2d 951, 958, 220 USPQ 592, 596, 598-99 (Fed. Cir. 1983), *cert. denied*, 469 U.S. 835, 225 USPQ 232 (1984)). Therefore, the question to be answered is: Does the claimed invention of a graphical representation of the at least one alternative splice variant determined using intensity values detected from probe sets disposed on probe arrays and the correlated annotation datum, have at least one specific and substantial asserted use?

Applicants respectfully assert that it most certainly does.

For example, the Examiner has indicated that claim 2 meets the requirements for utility. Claim 2 further limits the probe arrays of claim 1 to a specific diagnostic use. In other words, the Examiner has indicated that claim 2 further limits the result of claim 1 to one specific utility (i.e. a diagnostic utility), and thus specific utility for claim 1 is asserted. Such assertions of at least one use or objective to satisfy utility is consistent with the rulings in *Stiftung v Renishaw and Raytheon Co. v. Roper Corp.* (See *Stiftung v Renishaw PLC*, 945 F.2d 1173, 20 USPQ2d 1094 (Fed. Cir. 1991) and *Raytheon Co. v. Roper Corp.* referenced above).

Applicants also respectfully assert that the Examiner does acknowledge, but dismisses, an assertion of utility for probe arrays described in the application. In the Office Action mailed 2/11/2005 the Examiner states "such disclosed utility is not applicable to the instant claims" in reference to paragraph [0005] (In the Applicants

version paragraph [0005] begins on the 3rd page, line 17 (paragraph begins after with "In accordance with a particular embodiment...")) of the specification that states:

> "Also included in the method are the acts of correlating alternative splice variants with annotation data; and providing to the user, over a network, a graphical representation of the alternative splice variant and the correlated annotation data. The probe arrays may be constructed and arranged to diagnose a disease and/or medical condition, or for use in conducting research. Non-limiting examples of probe arrays constructed and arranged to diagnose a disease include probe arrays aimed at any one or more of the following applications: predisposition for disease or condition; screening; diagnosis; prognosis; pharmacogenomic applications (e.g., drug therapy selection and/or optimization), therapy selection and/or optimization for non-drug or combined therapies; monitoring of treatment response; and/or monitoring of disease progression, remission, and other indicators."

Applicants respectfully assert that the above disclosure clearly sets forth specific and substantial utility for the claimed probe arrays and a graphical display of alternative splice variants and correlated annotation data identified using the probe arrays. The Examiner has presented no arguments that the claimed probe arrays or graphical display are different than the disclosed probe arrays or graphical display, and thus there is no support for the Examiners assertion that the disclosed utility does not apply to the claims.

Along the same lines, Applicants also respectfully remind the Examiner that the bar for utility is not high where the invention is "useful" if it is capable of providing some identifiable benefit as discussed in *Juicy Whip Inc. v. Orange Bang Inc.* (See *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 51 USPQ2D 1700 (Fed. Cir. 1999)). Applicants respectfully assert that probe arrays and a graphical display of alternative splice variants and correlated annotation data identified using the probe arrays are described in the application and are also well known in the art as capable of analyzing

biological molecules such as nucleic acids, and that the results of said analysis clearly provide identifiable benefits to the public that include diagnostic and research benefits.

For the reasons described above, Applicants have shown that there is no requirement in the law for the claims to recite specific utility limitations. Applicants have further shown that specific and substantial utility for the claimed invention is asserted in both claim 2, the description in paragraph [0003], and further assert that it is known in the general state of knowledge of the art. If need be, Applicants will provide further examples of asserted utility but feel that it would be redundant to elaborate further because it is clear that utility is asserted in the examples already provided.

Therefore, Applicants respectfully assert that each of claims 1, 5, 19, 33, and 84 comply with 35 U.S.C. §101 and are thus patentable. Additionally, Applicants assert that each of claims 2, 4, 6-9, 11-16, 20-23, 25-30, 34-35 each depend from either claims 1, 5, 19, or 33 and are thus also patentable for the same reasons.

Reply to Claim Rejections – 35 U.S.C. §112, First Paragraph

Claims 1-5, 7, 11-21, 24-25, 28-37, and 43-45 are rejected under 35 U.S.C. §112 first paragraph.

Upon entry of the present amendments to the specification, Applicants respectfully assert that the invention is sufficiently described to enable one of ordinary skill in the related art to make and use the invention. In particular, Applicants have amended the description to include an example of a process of fitting hybridization intensity data and models of genomic structure to determine the presence of alternative splice variants. The example was described in US Provisional Patent Applications Serial

No. 60/398,958 which was properly incorporated by reference in it's entirety (see paragraph [0120]) in the present application as originally filed.

Applicants respectfully reiterate the assertion that the subject matter of the present amendments was previously incorporated by reference and that no new matter is presented by these amendments.

Therefore, Applicants respectfully request that the rejection be withdrawn.

Reply to Claim Rejections – 35 U.S.C. §112, Second Paragraph

Claims 1-5, 7, 11-21, 24-25, 28-37, and 43-45 are rejected under 35 U.S.C. §112 second paragraph.

With respect to the rejection of claims 1 and 43, Applicants respectfully disagree with the Examiner and assert that the limitations of intensity values detected from each probe set is definite. Applicants respectfully remind the Examiner that the current state of the law requires that claim language be evaluated in light of (1) the content of the particular application disclosure, (2) the claim interpretation that would be given by one of ordinary skill in the pertinent art at the time the invention was made, and (3) the teachings of the prior art. Applicants respectfully assert that those skilled in the art would appreciate the scope of the claimed limitations of "intensity values detected from each probe-set" in light of the specification (See *In re Wiggins*, 488 F.2d 538, 179 USPQ 421, 423-24 (C.C.P.A. 1973)).

For example, the specification describes a probe set as having one or more probes in paragraph [0080] and further describes generating for each probe a single value representative of the intensities of pixels measured by a scanner for that probe in

paragraph [0063]. Applicants respectfully assert that upon reading the description one of ordinary skill would understand that probe sets comprise one or more probes that are associated with intensity values detected from the probes.

However, in the interest of furthering prosecution Applicants have amended claims 1, 3, 28, and 43 to include the limitations of each probe set comprising one or more probes and the intensity values detected from the probes. Therefore Applicants respectfully assert that the amended limitations add further clarity to the claims and respectfully request that the rejections be withdrawn.

Applicants have also amended claims 1, 3, 28, and 43 to add clarity of the user selection comprising one or more probe set identifiers and one or more intensity values. Applicants respectfully assert that the amended limitations clarify the selection and respectfully assert are definite. Applicants respectfully request that the rejections be withdrawn.

With respect to the rejection of claims 1, 3, 28, and 43, Applicants respectfully assert that the term "fitting" is described in the application with respect to fitting data to models of genomic structure (see paragraphs [0119] and [0120]). Applicants also believe that further support is added by the present amendment to the description comprising an example of fitting intensity data to models of genomic structure. Applicants respectfully assert that one of ordinary skill in the related art would understand the limitations of fitting intensity data to a model of genomic structure in light of the description and amendments to the specification, and respectfully request that the rejections be withdrawn.

CONCLUSION

In conclusion, Applicants have amended the specification to include an example of fitting data to models of genomic structure. Applicants have also each of claims 1, 3, 28, and 43 to include add clarity to the claims. Applicants, therefore respectfully assert that each of the claims are patentable.

For these reasons, Applicants believe all pending claims are now in condition for allowance. If the Examiner has any questions pertaining to this application or feels that a telephone conference would in any way expedite the prosecution of the application, please do not hesitate to call the undersigned at (781) 280-1522.

The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account 01-0431.

Applicants respectfully request that a timely Notice of Allowance be issued in this case.

Respectfully submitted,

By 

William R. McCarthy III Reg. No.: 55,788

Attachments

Copy of US Provisional Patent Application Serial No. 60/398,958

Customer No.: 22886
Legal Department
Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051
Tel: 408/731-5000
Fax: 408/731-5392

1UZ 3508

UNITED STATES PATENT AND TRADEMARK OFFICE

COMMISSIONER FOR PATENTS
UNITED STATES PATENT AND TRADEMARK OFFICE
WASHINGTON, D.C. 20231
www.uspto.gov

APPLICATION NUMBER	FILING DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	DRAWINGS	TOT CLAIMS	IND CLAIMS
60/398,958	07/26/2002		160	3508	12		

CONFIRMATION NO. 9575

22886
AFFYMETRIX, INC
ATTN: CHIEF IP COUNSEL, LEGAL DEPT.
3380 CENTRAL EXPRESSWAY
SANTA CLARA, CA 95051

OIPE IAP52 APR 06 2006 PATENT & TRADEMARK OFFICE

FILING RECEIPT

OC000000008638036

Date Mailed: 08/16/2002

Receipt is acknowledged of this provisional Patent Application. It will not be examined for patentability and will become abandoned not later than twelve months after its filing date. Be sure to provide the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION when inquiring about this application. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please write to the Office of Initial Patent Examination's Filing Receipt Corrections, facsimile number 703-746-9195. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections (if appropriate).**

DOCKETED
AUG 21 2002
BY

Applicant(s)

Hui Wang, Santa Clara, CA;

If Required, Foreign Filing License Granted 08/15/2002

Projected Publication Date: Not Applicable

Non-Publication Request: No

Early Publication Request: No

Title

Method of analyzing alternative splicing

LICENSE FOR FOREIGN FILING UNDER
Title 35, United States Code, Section 184
Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Office of Export Administration, Department of Commerce (15 CFR 370.10 (j)); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

PATENT

Attorney Docket No. 3508

METHOD OF ANALYZING ALTERNATIVE SPLICING

Inventors: Hui Wang

Assignee:
Affymetrix, Inc.
a Corporation Organized under the laws of Delaware

Entity:
Large

Affymetrix, Inc.
Attn: Legal Department
3380 Central Expressway
Santa Clara, CA 95051
(408) 731-5000

METHOD OF ANALYZING ALTERNATIVE SPLICING

FIELD OF INVENTION

The present invention is related to biological data analysis methods and computer program products.

BACKGROUND OF THE INVENTION

There is a great need in the art for methods for analyzing splice variants.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a cartoon representation of alternative splicing process. The colored blocks represent exons. The thicker straight lines between blocks represent introns. The curved thin lines connecting blocks indicate that the sequences between these two exons are spliced out and these two exons are connected with each other. In this example, the gene contains 5 exons (a, b, c, d, e). It is first transcribed into pre-mRNA. Pre-mRNA then undergoes alternative splicing and 3 different variants are generated. Each variant have different combination of exons. These variants can potentially be translated into different proteins. Alternative splicing is one way to create protein diversities.

Figure 2 shows a process of outputing relative transcript concentrations by inputting gene structure information and hybridization intensity data through model fitting.

Figure 3 shows that once the relative transcript concentration is obtained by model fitting through inputting gene structure information and hybridzation intensity data, a gene expression profile can be created from the relative concentration.

Figure 4 shows the optimization process, data is first processed by generating an initial value, then the difference is repeated calculated until optimization.

Figure 5 shows that by inputting combined samples from a population and a model referencing a given genotype are inputted, allele frequency can then be determined through model fitting.

Figure 6 shows one possible arrangement of computer software modules to output reltive concentration. Gene structure information and hybridization are used as inputting modules.

Figure 7 illustrates a matrix representation of gene structures. This representation can answer questions such as which features are contained in a transcript, which transcript contains certain combination of features etc.

Figure 8 illustrates a matrix representation of probe intensities in a multiple experiments setup, where $E^{(j)}$ denotes the *j*th experiment, $P_{fn,i}$ is the *i*th probe of *fn*th feature. $y_{i,j}$ denotes the intensity value for *i*th probe in *j*th experiment.

Figure 9 illustrates the modeling result from spiked CD44 transcripts in yeast background with the CD44 exon and junction probes.

Figure 10 shows the changes of the sum of squared differences of observed intensities and predicated intensities for all the probes in all experiments, the fast convergence to a stable state indicate a good data fitting.

Figure 11 shows the CD44 modeling results in detail. Gene structure information and probes are listed on left. The graphs on top display the actual concentration and the predicted relative concentration from modeling. The blocks in the center plot the residuals for each experiment, high residual is indicated by a blue color and low residual by red. The initial value assigned to probe affinity is shown on the left, and the relative probe affinity term derived from model fitting is shown on the right.

Figure 12 lists the significance of CD44 expression level in tumorigenesis and metastasis.

SUMMARY OF THE INVENTION

In one aspect of the invention, methods are provided for determining relative concentrations of splice variants. In some embodiments, the methods include inputting the hybridization intensity; inputting gene structure information; subjecting said hybridization intensity and gene structure information to model fitting; and deriving relative concentration of splice variants and probe affinity terms.

In another aspect of the invention, methods are provided for creating a gene expression profile. In some embodiments, the methods include inputting expression hybridization intensity; inputting gene structure information; subjecting said expression hybridization intensity and gene structure information to model fitting; obtaining relative concentration of expressed gene; and creating an expression profile.

In yet another aspect of the invention, methods are provided for determining allele frequency. In some embodiments, the methods include inputting combined samples from a population; inputting a model referencing a given genotype; subjecting said combined samples and referenced genotype to model fitting; and deriving allele frequency of said genotype.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention has many preferred embodiments and relies on many patents, applications and other references for details known to those of the art. Therefore, when a patent, application, or other reference is cited or repeated below, it should be understood that it is incorporated by reference in its entirety for all purposes as well as for the proposition that is recited.

As used in this application, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "an agent" includes a plurality of agents, including mixtures thereof.

An individual is not limited to a human being but may also be other organisms including but not limited to mammals, plants, bacteria, or cells derived from any of the above.

Throughout this disclosure, various aspects of this invention can be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers

within that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example hereinbelow. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as Genome Analysis: A Laboratory Manual Series (Vols. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual (all from Cold Spring Harbor Laboratory Press), Stryer (anyone have the cite), Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London , all of which are herein incorporated in their entirety by reference for all purposes.

The practice of the present invention may also employ conventional computational biology methods, software or systems. Basic computational biology methods are described in, e.g., Setubal and Meidanis, et al., 1997, Introduction to Computational Molecular Biology, PWS Publishing Company, Boston; Human Genome Mapping Project Resource Centre (Cambridge), 1998, Guide to Human Genome Computing, 2nd Edition, Martin J. Bishop (Editor), Academic Press, San Diego;

Salzberg, Searles, Kasif, (Editors), 1998, Computational Methods in Molecular Biology, Elsevier, Amsterdam;

The present invention can employ solid substrates, including arrays in some preferred embodiments. Methods and techniques applicable to polymer (including protein) array synthesis have been described in U.S.S.N 09/536,841, WO 00/58516, U.S. Patents Nos. 5,143,854, 5,242,974, 5,252,743, 5,324,633, 5,384,261, 5,424,186, 5,451,683, 5,482,867, 5,491,074, 5,527,681, 5,550,215, 5,571,639, 5,578,832, 5,593,839, 5,599,695, 5,624,711, 5,631,734, 5,795,716, 5,831,070, 5,837,832, 5,856,101, 5,858,659, 5,936,324, 5,968,740, 5,974,164, 5,981,185, 5,981,956, 6,025,601, 6,033,860, 6,040,193, 6,090,555, and 6,136,269, in PCT Applications Nos. PCT/US99/00730 (International Publication Number WO 99/36760) and PCT/US 01/04285, and in U.S. Patent Applications Serial Nos. 09/501,099 and 09/122,216 which are all incorporated herein by reference in their entirety for all purposes.

Patents that describe synthesis techniques in specific embodiments include U.S. Patents Nos. 5,412,087, 6,147,205, 6,262,216, 6,310,189, 5,889,165, and 5,959,098. Nucleic acid arrays are described in many of the above patents, but the same techniques are applied to polypeptide arrays.

The present invention also contemplates many uses for polymers attached to solid substrates. These uses include gene expression monitoring, profiling, library screening, genotyping, and diagnostics. Gene expression monitoring, and profiling methods can be shown in U.S. Patents Nos. 5,800,992, 6,013,449, 6,020,135, 6,033,860, 6,040,138, 6,177,248 and 6,309,822. Genotyping and uses therefor are shown in USSN 10/013,598, and U.S. Patents Nos. 5,856,092, 6,300,063, 5,858,659, 6,284,460 and 6,333,179. Other

uses are embodied in U.S. Patents Nos. 5,871,928, 5,902,723, 6,045,996, 5,541,061, and 6,197,506.

The present invention also contemplates sample preparation methods in certain preferred embodiments. For example, see the patents in the gene expression, profiling, genotyping and other use patents above, as well as USSN 09/854,317, Wu and Wallace, Genomics 4, 560 (1989), Landegren et al., Science 241, 1077 (1988), Burg, U.S. Patent Nos. 5,437,990, 5,215,899, 5,466,586, 4,357,421, Gubler et al., 1985, Biochemica et Biophysica Acta, Displacement Synthesis of Globin Complementary DNA: Evidence for Sequence Amplification, transcription amplification, Kwoh et al., Proc. Natl. Acad. Sci. USA 86, 1173 (1989), Guatelli et al., Proc. Nat. Acad. Sci. USA, 87, 1874 (1990), WO 88/10315, WO 90/06995, and 6,361,947.

The present invention also contemplates detection of hybridization between ligands in certain preferred embodiments. See U.S. Pat. Nos. 5,143,854, 5,578,832; 5,631,734; 5,834,758; 5,936,324; 5,981,956; 6,025,601; 6,141,096; 6,185,030; 6,201,639; 6,218,803; and 6,225,625 and in PCT Application PCT/US99/ 06097 (published as WO99/47964), each of which also is hereby incorporated by reference in its entirety for all purposes.

The present invention may also make use of various computer program products and software for a variety of purposes, such as probe design, management of data, analysis, and instrument operation. See, U.S. Pat. Nos. 5,593,839, 5,795,716, 5,733,729, 5,974,164, 6,066,454, 6,090,555, 6,185,561, 6,188,783, 6,223,127, 6,229,911 and 6,308,170.

The present invention may also provide computer software and computer systems for performing the methods of the invention. Computer software products of the invention typically include computer readable medium having computer-executable instructions for performing the logic steps of the methods of the invention. Suitable computer readable medium include floppy disk, CD-ROM/DVD/DVD-ROM, hard-disk drive, flash memory, ROM/RAM, magnetic tapes and etc. The computer executable instructions may be written in any suitable computer language or combination of several languages.

Additionally, the present invention may have preferred embodiments that include methods for providing genetic information over the internet. See provisional application 60/349,546.

All cited references are incorporated herein by reference for all purposes.

I. Analyzing Alternative Splicing

Alternative splicing events are ubiquitous among eukaryotic organisms. Based on the recent studies, 30-60% of human genes undergo this process. In this process, instead of removing all the introns and generating one mRNA transcript, different combinations of exons can be alternatively spliced together to form many alternative splicing variants. Fig. 1 shows a cartoon representation of this process. In one embodiment of this invention, the relative transcript concentrations can be elucidated by inputting hybridization intensity and gene structure information. Fig. 2 shows a process of outputing relative transcript concentrations by inputting gene structure information and hybridization intensity data through model fitting. Once the relative concentration is calculated, a gene expression profile can be constructed. See Fig. 3.

II. Gene Structure and Hybridization Intensity

Microarray technology enables large scale, parallel monitoring of expression profiles of many genes. Microarray technology can be used to detect splice variants by measuring the intensity of the probe hybridized to gene features. In general, a gene is expressed differentially through alternative splicing of its transcript. A transcript may ~~include several gene features,~~ which refer to the sequences extracted from different splicing variants of the genes. A gene feature can be either exon, intron, or junction (exon-exon junction, exon-intron junction, intron-exon junction). Exon feature can be partitioned further depending on whether the exon is cassette exon or exons overlapping with others. Probes targetting gene features can be mapped to each of the transcript.

Gene structure includes all the transcripts of each gene and feature composition for each transcript. For example, a gene can have two transcripts A and B. Transcript A

includes 3 of the 5 features while transcript B has 4 features. The relationship between features and transcripts can be represented by matrix with values of 1s or 0s as shown in Fig. 7. The gene structure can be represented as follows:

$$\forall k, X_j^{(l)} = \sum_{k=1}^{m} g_{k,l} T_{k,j} \qquad (1)$$

Multiple probes are chosen to represent each gene feature. Though the probes have different properties, each measures the same concentration of a given transcript feature. By performing multiple experiments, the relationship of probes on a feature with experiments and intensities can be represented as a matrix in Fig. 8.

The probe intensity can be related to concentration and probe affinity using the following linear model:

$$y_{ij} = a_i x_j + \varepsilon_{ij} \qquad (2)$$

$$y_{ij} = a_i x_j + b_i + \varepsilon_{ij} \qquad (3)$$

Here, a_i is the affinity term for ith probe, b_i represents the background index for ith probe. The affinity term is probe-dependent. x_j is the relative concentration of feature in jth experiment. ε_{ij} denotes the error term, all those not explained by the other terms. Usually it is assumed to be normal distribution with mean 0 and variance σ^2. Formally, the error term can be written as $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Note the model is not limited in the form discussed above. Other model such as the models with multiplicative error proposed by Earl can also be used.

III. Model Fitting and Minimization

The above formulas can be rewritten as below for the *f(k)th* feature of a given gene:

$$y_{ij}^{f(k)} = a_i^{f(k)} x_j^{f(k)} + \varepsilon_{ij} \quad (4)$$

$$y_{ij}^{f(k)} = a_i^{f(k)} x_j^{f(k)} + b_i^{f(k)} + \varepsilon_{ij} \quad (5)$$

Combining with equation (1), we have for feature *f(k)* of a gene

$$y_{ij}^{f(k)} = a_i^{f(k)} \sum_{k=1}^{m} g_{k,f(k)} T_{k,j} + \varepsilon_{ij} \quad (6)$$

$$y_{ij}^{f(k)} = a_i^{f(k)} \sum_{k=1}^{m} g_{k,f(k)} T_{k,j} + b_i^{f(k)} + \varepsilon_{ij} \quad (7)$$

Differences between the predicated and observed intensity for each probe is minimized. A loss function is required for penalizing errors in predication. Many types of loss functions may be used for the same purpose, such as squared error loss function, absolute difference loss function. Here, the squared error loss function is applied to the model.

To minimize the squared difference between predicated and observed intensity value for all the probes of each gene (a set of features), the equations can be written as:

$$f(\overrightarrow{A},\overrightarrow{T}) = \sum_{k=1}^{nf}\sum_{j=1}^{ne}\sum_{i=1}^{np}(y_{ij}^{(k)} - a_i^{(k)} x_j^{(k)})^2 = \sum_{k=1}^{nf}\sum_{j=1}^{ne}\sum_{i=1}^{np}(y_{ij}^{(k)} - a_i^{(k)}(\sum_{k=1}^{m} g_{k,f(k)} T_{k,j}))^2 \quad (8)$$

$$f(\overrightarrow{A},\overrightarrow{T}) = \sum_{k=1}^{nf}\sum_{j=1}^{ne}\sum_{i=1}^{np}(y_{ij}^{(k)} - a_i^{(k)}x_j^{(k)} - b_i^{(k)})^2 = \sum_{k=1}^{nf}\sum_{j=1}^{ne}\sum_{i=1}^{np}(y_{ij}^{(k)} - a_i^{(k)}(\sum_{k=1}^{m} g_{k,f(k)}T_{k,j}) - b_i^{(k)})^2 \quad (9)$$

To minimize $f(\overrightarrow{A},\overrightarrow{T})$, some constraints or penalty terms are needed in order to solve it. The following constraints are added:

(10) $\sum_{i=1}^{np}(a_i^{(k)})^2 = constant$

(11) $a_i^{(k)} > 0$

(12) $T_{k,j}^{(k)} > 0$

Alternatively, the following penalty terms can be added to equations (7) and (8),

$$\gamma\sum_{k=1}^{nf}\sum_{i=1}^{np}(a_i^{(k)})^2$$

The solution is obtained by iteratively solving different sets of the parameters until convergence, yielding the relative concentration of each variant and the relative affinity term of each probe. One of ordinary skill in the arts can appreciate that a variety of other optimization methods may also be used, such as Maximum Likelihood and x^2. Fig. 4 shows the optimization process, data is first processed by generating an initial value, then the difference is repeated calculated until optimization.

IV. Creating Gene Expression Profiles

A gene expression profile may be created by inputting expression hybridization intensity and gene structure information. Through the model fitting described above, the relative concentration of expressed gene can be determined; an expression profile can be

created based on the relative concentration. Fig. 3 illustrates the creation of an expression profile.

V. Analyzing Allele Frequency

In one embodiment of this invention, allele frequency can be analyzed. In the situation of multiple SNPs from a mixture of different individuals, where the number of patterns present are given, the relative frequency of each pattern can also be calculated. As shwon in Fig. 5, combined samples from a population and a model referencing a given genotype are inputted. Allele frequency can then be derived through the model fitting discussed above.

VI. Software and System for Analysis

In one aspect of the invention, software products and systems are provided for performing the methods of the invention.

Gene structure information and hybridization are used as inputting modules in a computer software. Fig. 6 shows one possible arrangement of these modules to output reltive concentration.

VII. Example

This example illustrates one embodiment of methods of the invention and its application in determining relative concentration of splice variants and probe affinity.

A. Introduction

This is a general model for determining the relative concentration of different splice variants. This model dealt with intensities in probe level across multiple experiments (>=2 chips). The model took the gene structure and probe intensities as input data, and output the relative concentration of each variant and the affinity term of each probe. The probe intensities were initially assigned an arbitrary value, and the model would output a relative affinity term. These relative affinity terms could then be used to measure the quality of the probes. Probes with low affinity terms could be identified and replaced; the data could be refitted iteratively by using higher affinity probe to produce better results.

B. Experimental Protocols

A gene is expressed differentially through alternative splicing of its transcript. A transcript may include several gene features. Probe targeted to these features could be mapped to each of the transcript. The relationship of genes, transcript , gene features could be represented by Fig. 7.

Gene structure includes all the transcripts of each gene and feature composition for each transcript. For example, a gene could have two transcripts A and B. Transcript A included 3 of the 5 features while transcript B had 4 features. The relationship between features and transcripts could be represented by matrix with values of 1s or 0s described as follows:

Let G be an m by n matrix. m is the number of transcripts while n represents the number of features for a gene. The column $F^{(l)}$ denotes feature l, $T_{j,k}$ denotes kth transcript, it is also used to denote the concentration of kth transcript in experiment j. $g_{k,l}$

is the element of this matrix for kth transcript and lth feature, its value is either 1 or 0. See Fig. 7.

The matrix could be written using the following equation, where $X_j^{(l)}$ denoted the concentration measured by lth feature in experiment j. $X_j^{(l)}$ could be written as:

$$\forall k, X_j^{(l)} = \sum_{k=1}^{m} g_{k,l} T_{k,j} \qquad (1)$$

where $g_{k,l}$ is either 0 or 1. Equation (1) therefore represented the gene structure.

To model the data, multiple probes were used to represent each feature. These probes had different properties, however they measured the same concentration of a given transcript feature. In a setup of multiple experiments, the relationship of probes on a feature with experiments and intensities could be expressed as the matrix shown in Fig. 8.

A simple model was adopted to express the relationship between probes properties, concentrations and intensity measurements:

$$y_{ij} = a_i x_j + \varepsilon_{ij} \qquad (2)$$

$$y_{ij} = a_i x_j + b_i + \varepsilon_{ij} \qquad (3)$$

In the above equations, a_i was the affinity term for ith probe (which is arbitrarily assigned), b_i represented the background index for ith probe. This term was probe-dependent, x_j was the relative concentration of feature in jth experiment, and ε_{ij} denoted the error term. The error term included all factors not explained by the other terms, usually it is assumed to be normal distribution with mean 0 and variance σ^2. Formally, this could be written as $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The above formulas could be rewritten as follows for the *f(k)th* feature of a given gene:

$$y_{ij}^{f(k)} = a_i^{f(k)} x_j^{f(k)} + \varepsilon_{ij} \quad (4)$$

$$y_{ij}^{f(k)} = a_i^{f(k)} x_j^{f(k)} + b_i^{f(k)} + \varepsilon_{ij} \quad (5)$$

Combining these equations with equation (1), we have for feature *f(k)* of a gene:

$$y_{ij}^{f(k)} = a_i^{f(k)} \sum_{k=1}^{m} g_{k,f(k)} T_{k,j} + \varepsilon_{ij} \quad (6)$$

$$y_{ij}^{f(k)} = a_i^{f(k)} \sum_{k=1}^{m} g_{k,f(k)} T_{k,j} + b_i^{f(k)} + \varepsilon_{ij} \quad (7)$$

Differences between the predicated and observed intensity for each probe was minimized. A loss function was required for penalizing errors in predication. Many types of loss functions may be used for the same purpose, such as squared error loss function, absolute difference loss function. Here, the squared error loss function was applied to the model.

To minimize the squared difference between predicated and observed intensity value for all the probes of each gene (a set of features), the equations could be written as:

$$f(\overrightarrow{A},\overrightarrow{T}) = \sum_{k=1}^{nf} \sum_{j=1}^{ne} \sum_{i=1}^{np} (y_{ij}^{(k)} - a_i^{(k)} x_j^{(k)})^2 = \sum_{k=1}^{nf} \sum_{j=1}^{ne} \sum_{i=1}^{np} (y_{ij}^{(k)} - a_i^{(k)} (\sum_{k=1}^{m} g_{k,f(k)} T_{k,j}))^2 \quad (8)$$

$$f(\overrightarrow{A},\overrightarrow{T}) = \sum_{k=1}^{nf} \sum_{j=1}^{ne} \sum_{i=1}^{np} (y_{ij}^{(k)} - a_i^{(k)} x_j^{(k)} - b_i^{(k)})^2 = \sum_{k=1}^{nf} \sum_{j=1}^{ne} \sum_{i=1}^{np} (y_{ij}^{(k)} - a_i^{(k)} (\sum_{k=1}^{m} g_{k,f(k)} T_{k,j}) - b_i^{(k)})^2 \quad (9)$$

To minimize $f(\vec{A},\vec{T})$, some constraints or penalty terms were needed in order to solve the equations. The following constraints were added:

$$(10)\ \sum_{i=1}^{np}(a_i^{(k)})^2 = constant$$

$$(11)\ a_i^{(k)} > 0$$

$$(12)\ T_{k,j}^{(k)} > 0$$

Alternatively, the following penalty terms could be added to equations (7) and (8),

$$\gamma\sum_{k=1}^{nf}\sum_{i=1}^{np}(a_i^{(k)})^2$$

The solution was obtained by iteratively solving different sets of the parameters until convergence, yielding the relative concentration of each variant and the relative affinity term of each probe.

C. Result

This example demonstrates a general model that could be used to analyze alternative splicing. Spiked CD44 transcripts in yeast background was performed, and modeling results using the CD44 exon and junction probes are presented in Fig. 9. A near diagonal line (45 degrees) indicates good data prediction. The quality of the data fitting can also be examined by the residual. Fig. 10 shows the changes of the sum of squared differences of observed intensities and predicated intensities for all the probes in all experiments, the fast convergence to a stable state indicate a good data fitting. Fig. 11 shows the CD44 modeling results in detail. Gene structure information and probes are listed on left. The

graphs on top display the actual concentration and the predicted relative concentration from modeling. In, Fig. 11, the blocks in the center plot the residuals for each experiment, high residual is indicated by a blue color and low residual by red. After modeling, the residuals are noticeably lower.

In addition to relative splice variant concentration, relative probe affinity terms outputted can also be useful in improving data fitting. An illustration of this process is shown in Fig. 11. An initial arbitrary affinity term assigned to the probes yields a relative affinity term through the model. Probes with low affinity terms can then be discarded and the data refitted iteratively. In Fig. 11, the probe targeting exon 3 feature 1 should be discarded. Understanding CD44 differential expression patterns may lead to valuable insights regarding tumorigenesis and metastasis. See Fig. 12.

This example illustrates alternative splicing typing—when the possible splice variants are known in a given sample and relative concentration is desired. In addition, one of ordinary skill in the arts can also appreciate the discovery of new transcripts using this invention.

We claim:

1. A method for determining relative concentrations of splice variants comprising: Inputting the hybridization intensity; inputting gene structure information; subjecting said hybridization intensity and gene structure information to model fitting; and deriving relative concentration of splice variants and probe affinity terms.

2. A method for creating a gene expression profile comprising: Inputting expression hybridization intensity; inputting gene structure information; subjecting said expression hybridization intensity and gene structure information to model fitting; obtaining relative concentration of expressed gene; and creating an expression profile.

3. A method for determining allele frequency comprising: Inputting combined samples from a population; inputting a model referencing a given genotype; subjecting said combined samples and referenced genotype to model fitting; and deriving allele frequency of said genotype.